DOCUMENT RESUME

ED 249 229 TM 840 439

AUTHOR Johnson, Sylvia T.
TITLE The Test, the Tested, and the Test-Taking: A Model to Better Understand Test Performance.
PUB DATE Apr 84
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Models; *Performance Factors; Student Characteristics; Testing; Testing Problems; *Test Results; Tests

ABSTRACT

A model is presented in which a test score is influenced by three factors: (1) the test itself, including the nature of the item tryout samples, the nature of the standardization group, the item scaling model and methods, the manner of item subset generation, logical strategies in test building, the format of questions, the test appearance, item content and features, and the choice of criterion; (2) characteristics of those taking the test, including race and racism, sex and sexism, motivation, socioeconomic status, school experiences, home experiences, ways of thinking, and interests and preferences; and (3) the setting and administrative procedures used in the testing, including personal characteristics of the tester, expectancies and beliefs of the tester, the physical surroundings, the mode of test preparation, and the timing or pacing of presentation. These factors affecting the test score may interact in complex ways. Part of the influence on test score is intentional and valid, and attempts are made to control other aspects. Many aspects, however, cannot be controlled. Yet, all need to be recognized as potential components of the score obtained. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED249229

THE TEST, THE TESTED, AND THE TEST-TAKING: A MODEL TO BETTER UNDERSTAND TEST PERFORMANCE

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sylvia T. Johnson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Sylvia T. Johnson
Howard University

TM 840 435

A paper presented at the annual meeting of the American Educational Research Assn., New Orleans, La., April, 1984

THE TEST, THE TESTED, AND THE TEST-TAKING: A MODEL TO BETTER UNDERSTAND TEST PERFORMANCE

Everyone here is familiar with the theme of my presentation here this morning. It is simply this: a test score is influenced by the measure itself; those people who are being tested, and the setting in which this activity occurs. The score becomes a product of the test itself, including its content, norming procedures, and construction; a product of the background characteristics and typical ways of thinking of those being tested; and a product of the characteristics of the setting and administrative procedures used in the testing.

The reason for still drawing attention to such a model at this point is that despite all that we know about factors influencing test scores, our understanding of what we are measuring with our test items has not increased substantially, and in many cases the practical process of actually writing test items is little changed, with the exception of the use of item-generating rules in criterion-referenced tests.

After a test is developed and placed in use, the resulting test score is generally interpreted as an adequate measure of the targeted construct. It's not that we don't know what factors can or do influence test scores -- its more that we are not sure what to do about it, or more precisely, what to do in a variety of given cases.

Surely the major development in measurement practice in recent years is the active use of procedures based on item response theory (IRT) in the actual construction of tests. This theoretical base has been widely available at least since the 1968 publication the Lord and Novick text, Statistical Theories of Mental Test Scores, which presented an introductory chapter on

BEST COPY

latent traits and item characteristic functions, followed by a set of chapters by Allan Birnbaum on latent trait models and information structures of tests, and the use of these concepts in classification and estimation of abilities. (Lord and Novick, 1968). The more recent work of Lord and many others has made this theoretical base more accessible to the practitioner (Lord, 1980). This process has also been aided by the generally increasing statistical sophistication of graduate programs in measurement.

The use of IRT models in test construction offers little information on what is actually being measured. We may use IRT models to calibrate test items to a common scale, and in so doing, we build within-test homogeneity and unidimensionality, and ease comparisons across tests and groups. To the extent that these outcomes increase the precision of the measurement, the quality of the measuring process is improved, but the score remains the product of the same basic set of factors. The recognized improvements that we have made in constructing tests by using item response theory models and methods does not impact on systematic sources of variance that may tend to rather uniformly affect subgroups, and thus alter the construct validity of the measure. If a systematic source of variance is pervasive across items, and is correlated with the "underlying" trait being measured, it may simply strengthen the assertation of unidimensionality, and be quite resistant to detection by statistical bias techniques.

For example, when the Department of Defense renormed existing test forms of the Armed Forces Qualification Test Battery, they administered their items to a carefully selected sample of young people nationally, with broad representation of racial minorities and of educational and income groups. A thoroughly done bias study, perhaps the most extensive ever done, showed little ethnic bias in the items. However, since the variability in

BEST COPY AVAILABLE

backgrounds of the testees was so great, a trio similar to the following would likely be found in the 23 year-old sample: a graduate student in English literature; a factory foreman who performed well in earning a high school diploma has a wife and two children, and has not taken a test in the past five years; and a young man who left high school at age 16, after about three years of marginal attendance and even poorer performance. If the test performance of these young people and others more or less like them on the same test items, is significantly different on a large proportion of the test items for any reasons other than ability, those differences are likely to be undistinguishable from the ability dimension. If differences occured on only a few items, this performance would be more distinguishable by item bias methods. But when performance differences are broadly systematic, they easily become indistinguishable from the construct of interest. Bias techniques are best at locating performance different from the usual pattern, regardless of whether typical differences are valid differences. The assignment of the three not so hypothetical young men to one of five different "mental categories" by the Department of Defense may not be so much a matter of test bias, but rather, the lack of recognition that the same set of test items may be measuring something quite different for each of these individuals. Sources of variance that raise or lower score of certain subgroups, and have the opposite effect on the majority group or on other subgroups, without substantially affecting the weighted mean of the total group are dealt with reasonably effectively by IRT methods. A wide variety of useful procedures have been developed for spotting items with response patterns divergent from a large body of items being studied.

However, if a source of variance operates in such a way that the patterns are not divergent, or if the direction of the effect is to systematically reduce scores of subgroups and lower the weighted mean of the total group, then the source cannot be identified by IRT methods. In fact, it may serve to strengthen the unidimensionality determination made for a large set of items tried out over a large number of subjects. For example, suppose that instructional quality, or SES level within school districts is related to achievement, then a large scale study cutting across many school districts may incorporate this systematic factor into the achievement dimension.

This is not to say that the creatively growing set of IRT procedures ought not be used to generally develop and improve tests, but to simply remind us that their use does not negate, in fact it strengthens the need to carefully study the ways in which individual and group differences can affect test performance, and to recognize these factors in constructing tests, in more precisely defining the construct which we are measuring, and perhaps most of all, interpreting the meaning of the obtained score. This is not an easy task, because correlational procedures for construct validation, including factor analysis, are affected by the reliability of the possible variance sources examined; and ability differences are generally more reliable than such fuzzy sources as background characteristics and styles or modes of thinking.

One may have information which suggests that a particular background factor has a systematic effect on reading test performance, but is it a constant factor for everyone with that background? Does it interact with other factors, including characteristics of the test items? Are there ways to estimate the effect? Can we use this information in both instrument

construction and educational practice? These issues deserve extended study, and the need for such study is increased rather than decreased as the arsenal of measurement technology increases in sophistication.

Another major development perhaps is not as generally recognized for its growing impact on test development as is IRT Theory, but it has more potential for answering some of the questions noted above. It is the rapidly expanding area in cognitive psychology that has to do with what sort of process the individual is really undertaking when a given task or problem is completed or solved. The conceptualization, definition, and the qualitative and quantitative appraisal of these processes is the focus of the work in this area. A recent article by Sternberg (1984) offers a further developed theoretical framework for work in the area.

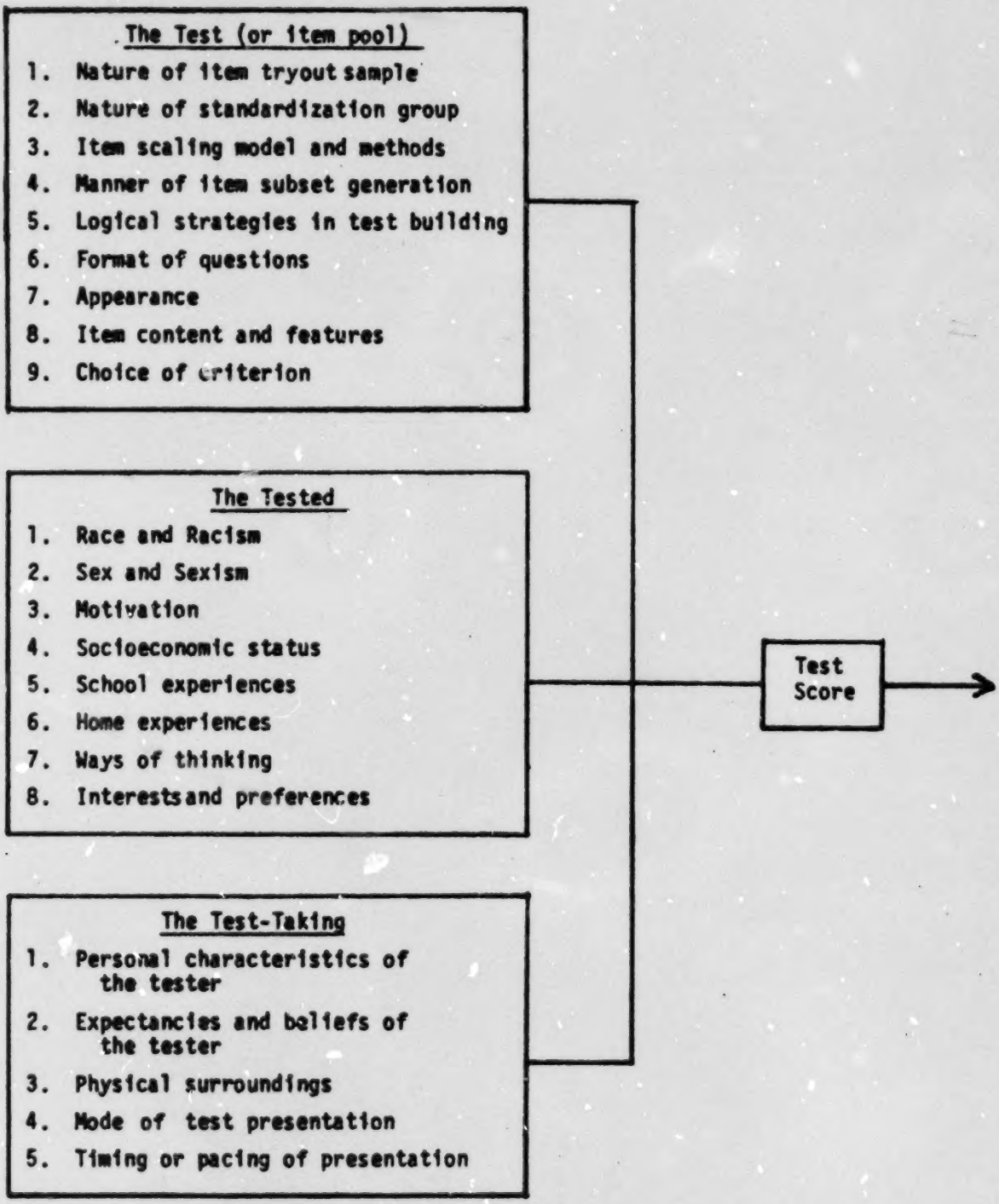
The fact that educational measurement is a hybrid field, and involves the understanding in depth of more than one discipline is both exciting and frustrating. But measuring characteristics of people does involve understanding the people, the measures, and under what circumstances the two meet.

Thus, it is increasingly necessary to examine the total context for drawing the meaning of the test score, and for this reason the model is proposed. The model is not all inclusive or exhaustive, but simply suggests some areas within the total testing process that may impinge on test performance. It is adapted from a model of sources of test bias, which appeared in an earlier publication (Johnson, 1979). The changed title reflects an emphasis on what elements do go into a test score, without necessarily placing a value on their validity.

The model is presented in figure 1.

Figure 1

Factors Affecting Test Score


The Test (or item pool)
1. Nature of item tryout sample
2. Nature of standardization group
3. Item scaling model and methods
4. Manner of item subset generation
5. Logical strategies in test building
6. Format of questions
7. Appearance
8. Item content and features
9. Choice of criterion
The Tested
1. Race and Racism
2. Sex and Sexism
3. Motivation
4. Socioeconomic status
5. School experiences
6. Home experiences
7. Ways of thinking
8. Interests and preferences
The Test-Taking
1. Personal characteristics of the tester
2. Expectancies and beliefs of the tester
3. Physical surroundings
4. Mode of test presentation
5. Timing or pacing of presentation
Test Score

BEST COPY

Most of the elements of the model have been discussed extensively elsewhere, and need little elaboration here. The basic elements relate to the process of test development and item and test construction, a range of environmental and biological characteristics, and aspects of the setting in while the testing occurs.

Since "the test" may be a randomly selected group of items, a chunk from a matrix, or another form of subset of test items, the test-related factors also refer to that larger set of items. The test factors include the nature of the group on whom the items are first tried out as well as the standardization group; the item scale model and methods employed, the logical strategies used in test building, the format of test questions, the appearance of the test, the content and special features of test items, and the choice of criterion for validation.

The factors affecting test score that arise from aspects relating to those who are tested include two important biosocial attributes, race and sex. Since the meaning of these factors arises out of interaction of individuals with positive and negative factors of the environment, the characteristic and the negative reaction must both be considered. Thus, these factors are labelled race and racism, and sex and sexism. Other factors relating to those who are tested include motivation, socioeconomic status, school experiences, home experiences, characteristic ways of thinking, and interests and preferences.

The aspects of the "test-taking" which affect test score include personal characteristics and beliefs of the tester; physical surroundings; the mode of test presentation, which may be the typical paper and pencil exam, an exam presented by computer, or an exam presented and modified by computer according to answers given to presented items; and the timing or pacing of the test presentation.

BEST COPY

Clearly, these factors affecting the test score may interact in complex ways. Part of the "influence" on test score is certainly intentional and valid, and other aspects we try to control. Many aspects cannot be controlled. Yet, all need to be recognized as potential components of the score obtained.

The empirical studies reported today examine a variety of test factors and background characteristics in terms of traditional school performance as well as performance on non-traditional measures. The populations include inner-city black public primary school children, older black youngsters from varied background, and pre-adolescent children in parochial schools. In addition, one paper reviews research and measurement issues in the appraisal of basic skills. Together, this work can suggest to us the broad and relatively untouched ground of non-traditional interpretations of scores on traditional instruments, as well as the use of non-traditional instruments and items to measure constructs of usual interest in a school setting. Our thinking in these areas must continue to expand along with, and hopefully ahead of our technology, if we want to be fair to all those who interact with us in the measuring process.

BEST COPY

Bibliography

Johnson, S. T. The Measurement Mystique. Washington, D.C.: Institute for the Study of Educational Policy, Howard University, 1979.

Lord, F. M. Application of Item Response Theory to Practical Testing Problems. Hillside, N.J.: Erlbaum, 1980.

Lord, F. M. and Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass.; Addison-Wesley, 1968.